Pattern Classification with Memristive Crossbar Circuits

Dmitri B. Strukov
Department of Electrical and Computer Engineering Department
UC Santa Barbara
Santa Barbara, CA, USA, 93106-9560

Abstract: *Neuromorphic pattern classifiers were implemented, for the first time, using transistor-free integrated crossbar circuits with bilayer metal-oxide memristors. 10×6- and 10×8-crosspoint neuromorphic networks were trained in-situ using a Manhattan-Rule algorithm to separate a set of 3×3 binary images: into 3 classes using the batch-mode training, and into 4 classes using the stochastic-mode training, respectively. Simulation of much larger, multilayer neural network classifiers based on such technology has shown that their fidelity may be on a par with the state-of-the-art results.*

Keywords: Memristor; resistive switching; RRAM; artificial neural networks; perceptron; pattern classification; deep learning; convolutional neural network networks.

Introduction

Deep-learning convolutional neural networks (DLCNN), which are essentially multilayer perceptrons (MLP) with restricted connectivity between some layers (Fig. 1a), have been demonstrated to achieve some of the best classification performances on a variety of benchmark tasks [1]. The major challenge in building fast and energy-efficient networks of this type in hardware is performing efficient vector-by-matrix multiplication, which in turn requires compact implementation of synaptic weights [2]. CrossNet circuits have emerged as an efficient solution to these challenges [2]. In such a network, neural cell bodies are mimicked with analog CMOS circuits, which communicate via passive crossbars with integrated tunable resistive devices ("memristors") [3-6], playing the role of synapses [7-12] (Figs. 1b-e). Main goals of this work were to demonstrate the first neural networks with integrated crossbar circuits, and evaluate possible performance of larger classifiers based on this emerging technology.

Experimental Results

A 12×12 crossbar with 200-nm lines separated by 400-nm gaps (Fig. 2a), with a $Pt/Al_2O_3/TiO_{2-x}/Ti/Pt$ memristor at each crosspoint, was fabricated using a standard lift-off patterning. The Al_2O_3/TiO_{2-x} stack was deposited by reactive sputtering, with titanium oxide stoichiometry controlled precisely via the oxygen flow control. The thickness and stoichiometry were optimized to achieve low forming voltages (<2 V) and highly nonlinear *I-V* curves with a ~10 ratio of current values at the switching voltage (~1.5 V) and at a half of it (Fig. 2b). The most outstanding feature of such memristors is their low variability (Fig. 2c); together with nonlinear *I-V* and low forming voltages it has enabled forming of most of the devices in crossbar array. Other important characteristics are the ~100 ON/OFF current ratio at ~0.1 V, a switching endurance of at least 5,000 cycles, an estimated retention of at least 10 years at room temperature, and operation currents between ~100 nA and ~100 μA [9, 10]. Using short (e.g., 500 μS) pulses makes both set and reset switching processes fairly continuous, enabling gradual tuning of device conductance with an at least 5-bit precision [10] even using a very simple (suboptimal) feedback algorithm [11]. Such precision is already acceptable for some neural network applications [2, 12].

During classifier's operation (Figs. 1e, 3a, 4a), the vector-by-matrix multiplication of the input signals (represented with voltages) by weights (represented by memristor conductances) is performed on the physical level, in analog domain, using Ohm's and Kirchhoff's laws, by applying the input voltages to crossbar's row lines and reading out the currents flowing into virtually grounded column lines (Fig. 1e, 4a). The training was performed in-situ in both the batch and stochastic modes, using the Manhattan-Rule algorithm (Fig. 3) [13]. This rule is convenient for crossbar

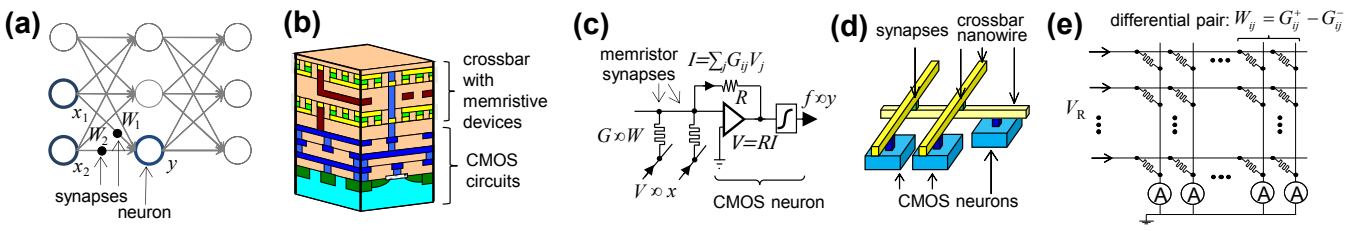

Figure 1. Neuromorphic network implementation with CrossNet circuits [2]: (a) A graph representation of a multilayer perceptron; (b) a cartoon of a hybrid CMOS/memristor (CMOL) integrated circuit; (c) analog implementation of the dot-product, (f) its mapping on the hybrid circuit, and (e) the implementation of vector-by-matrix multiplication using a memristive crossbar. (It shows that if negative weight values are required, a synapse may be implemented as a pair of memristors.)

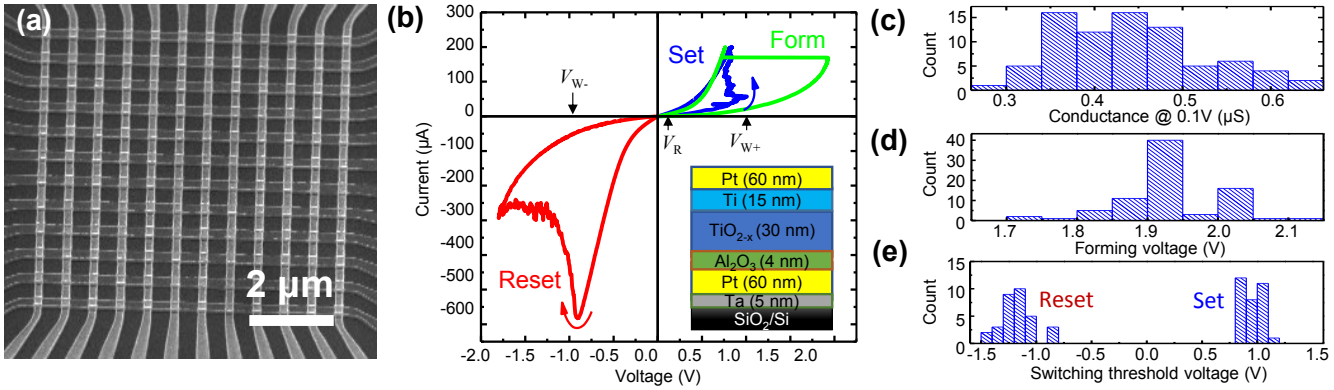

Figure 2. Crossbar circuit with integrated Al_2O_3/TiO_{2-x} resistive switching devices: (a) micrograph of a 12×12-crosspoint crossbar; (b) typical quasi-dc *I-V* curves of memristor forming and switching, with the inset showing the device stack; and histograms of: (c) conductances before forming, (d) forming voltages, and (e) effective switching threshold voltages. (The threshold is conditionally defined as the point at which device's resistance is changed by at least 2 kΩ upon application of a 500-μs voltage pulse train with a slowly increasing amplitude, starting from high/low conductive state for reset/set transitions.)

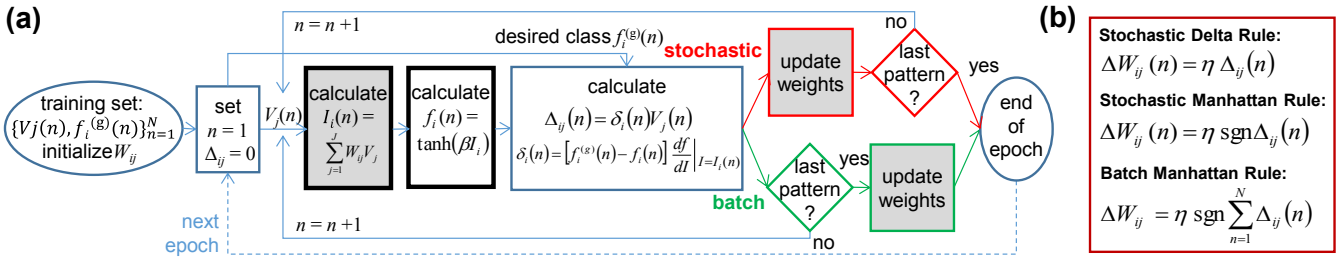

Figure 3. In-situ training of a single-layer perceptron classifier: (a) flow chart of one epoch for batch- and stochastic-mode training algorithms. Gray-shaded boxes show the steps implemented inside the crossbar, while those with solid black borders denote the only steps required for performing the feedforward (classification) operation.

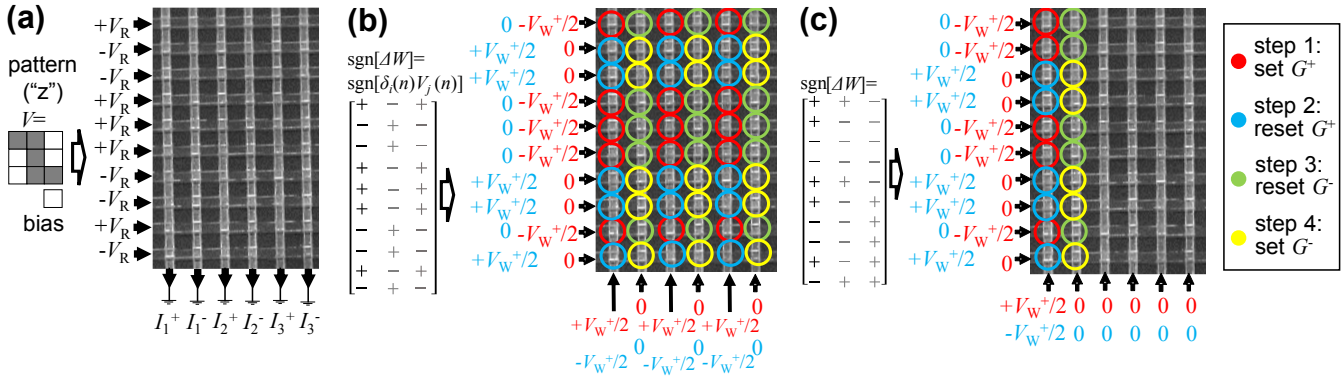

Figure 4. Physical-level description of the classification experiment: (a) example of operation of classifier using a 10×6 fragment of the crossbar; example of weight adjustment for (b) stochastic and (c) batch training for a specific error matrix. Panels (b) and (c) show the voltages only for first two steps. The read and write biases were always V_R = 0.1 V and $V_W^{\pm}$ = ±1.3 V, respectively (Fig. 2b).

circuit implementation, due to the use of only the sign information of the conventional Delta-Rule algorithm's result. The advantage of stochastic training is that the weight update for the whole crossbar (of any size) may be performed in just four steps by applying pulses in parallel to rows and columns of the crossbar (Fig. 4b) [12]. Namely, the weights are grouped into four sets, each corresponding to a particular combinations of signs of $V(n)$ and $\delta(n)$, and the weight in each group are updated in parallel. On the contrary, in the batch mode the weights in different columns (or rows) have to be updated sequentially (Fig. 4c), so that the update time grows linearly with crossbar size. Additionally, the batch mode training may come with a large area overhead when implemented on-chip, due to the need of computing and storing intermediate results for the weight update [17].

Generally, device-to-device variations of the switching threshold present a significant challenge for the in-situ training, because exponential switching dynamics [6, 11] amplifies even slight threshold variations. Additionally, the change in conductance depends on the initial conductance of the device. In this context, the fact that we have been able to achieve successful convergence for both the batch and stochastic in-situ training, even despite substantial device-to-device variations in switching dynamics, is highly encouraging (Fig. 5). The batch-mode training gave more stable convergence, while the update dynamics for that stochastic training was very close to that in the software-implemented network [10].

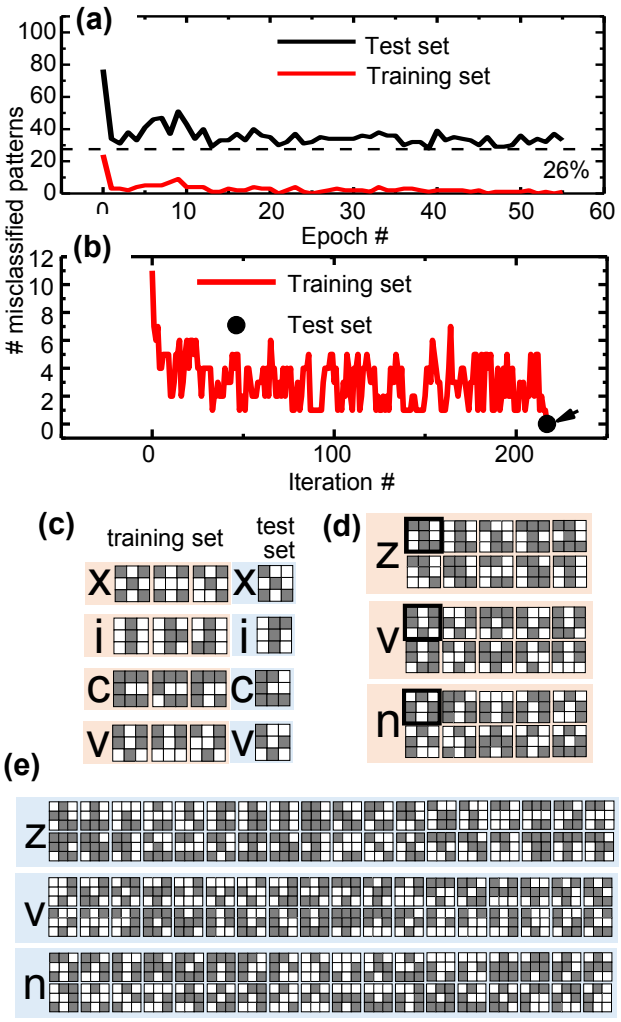

Figure 5. Results of pattern classification experiments: the convergence of network's output in the process of in-situ training for the (a) batch and (b) stochastic training modes; (c-e): the training and test images used for (c) batch and (d, e) stochastic training experiments. For the batch training, one epoch is the input of 30 patterns, while for stochastic training, one iteration is the application of one pattern. The batch mode (d) training / (e) test images are formed by flipping one pixel / two pixels of the "ideal patterns" shown with the solid border.

Simulations Results

In another part of this work, an accurate, data-verified model of adaptation of our memristors [14] was used to simulate the performance of pattern classifiers, based on a large-scale fully connected MLP and DLCNN, on several representative benchmarks [15], using both in-situ and ex-situ training [16, 17]. Similarly to the experimental results, the classification performance was worse for the stochastic Manhattan-Rule training (Table 1a). However, a simple "variable-amplitude" variation of the training scheme [16, 17] allows an implementation of the more efficient Delta-Rule algorithm (Fig. 3b), which dramatically improves the stochastic-mode fidelity and achieves state-of-the-art performance for the batch training (Table 1a-b). In such variable-amplitude scheme, write voltages proportional to $\log[V(n)]$ and $\log[\delta(n)]$, of specific polarity, are applied to the corresponding lines of the crossbar. Since the change of device conductance is roughly exponential in the applied voltage, this procedure results in weight update proportional to the product of δV, thus implementing the Delta Rule directly in the crossbar, without the need of its calculation in external hardware. The simulation results also show that the in-situ training is inherently robust to various network defects (Fig. 6), and that an 8-bit weight import at ex-situ training is sufficient to avoid classification fidelity degradation [17].

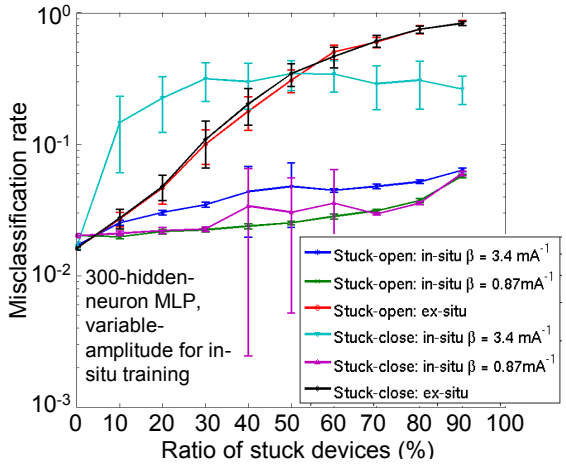

Figure 6. MNIST dataset classification fidelity of a MLP as a function of the fraction of stuck-on-open or stuck-on-close devices, for several training approaches.

Discussion and Summary

We have experimentally demonstrated an artificial neural network using memristors integrated into a dense, transistor-free crossbar circuit. We believe that this demonstration is a significant step toward analog-hardware implementation of practical artificial neural networks. The simulation of such scaled-up networks, using a quantitatively verified model of our memristors, has shown that their performance can be competitive to the state-of-the-art software implementations. Moreover, recent experiments [18] with similar but smaller (so far, discrete)

Table 1. Classification fidelity for (a) 300-hidden-neuron MLP network tested on the MNIST benchmark, and (b) DLCNN, with architectures similar to those in [15], tested on three indicated benchmarks. 500 patterns per batch were used for batch training.

(a)

Training mode	Software		Xbar in-situ				Xbar ex-situ 2% import	
			Manhattan Rule		Var. ampl.			
	best	average	best	average	best	average	best	average
Batch	-	-	1.98	2.06±0.09	1.47	1.62±0.07	-	-
Stochastic	1.57	1.75±0.07	19.26	20.16±1.3	4.06	4.31±0.33	1.54	1.62±0.04

(b)

Data set	Software		Xbar in-situ (var. ampl.)		Xbar ex-situ 2%		Xbar ex-situ 0.2%	
	best	average	best	average	best	average	best	average
MNIST	0.40	0.47±0.05	0.4	0.48±0.024	0.61	0.89±0.22	0.41	0.42±0.01
GTSRB	1.36	1.53±0.18	1.26	1.56±0.27	1.42	1.56±01	1.46	1.47±0.01
CIFAR10	15.63	15.91±0.22	15.67	15.87±0.22	19.77	20.29±0.43	15.5	15.8±0.01

memristors give hope that the metal-oxide memristor networks may be scaled down to at least 30-nm devices. According to theoretical estimates [2], such networks would enable CrossNets with an areal density higher than that of the human cerebral cortex, operating at much higher speed and with comparable energy efficiency.

Acknowledgements

This work was supported by AFOSR under MURI grant FA9550-12-1-0038, by DARPA under contract HR0011-13-C-0051UPSIDE via BAE Systems, Inc., and by the DENSO CORP., Japan. The author would like to acknowledge contribution by M. Prezioso, B. Hoskins, G. Adam, F. Merrikh-Bayat, I. Kataeva, and K.K. Likharev.

References

1. Krizhevsky, A., I. Sutskever, and G.E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in: *Proceedings of NIPS'12*, Lake Tahoe, NV, Dec. 2012, pp. 1097-1105.
2. Likharev, K.K., "CrossNets: Neuromorphic Hybrid CMOS/Nanoelectronic Networks," *Science of Advanced Materials*, Vol. 3, no.3, pp. 322–331, May 2011.
3. Waser, R., R. Dittman, G. Staikov, and K. Szot, "Redox-Based Resistive Switching Memories," *Advanced Materials*, Vol. 21, pp. 2632–2663, 2009.
4. Wong, H.S.P. et al., "Metal-Oxide RRAM," *Proceedings of IEEE*, Vol. 100, pp. 1951-1970, 2012.
5. Lu, W., D.S. Jeong, M. Kozicki, and R. Waser, "Electrochemical Metallization Cells – Blending Nanoionics into Nanoelectronics," *MRS Bulletin*, Vol. 37, no. 2, pp. 124-130, 2012.
6. Yang, J.J., D.B. Strukov, and D.R. Stewart, "Memristive Devices for Computing", *Nature Nanotechnology*, Vol. 8, pp. 13-24, Jan. 2013.
7. Yu, S. et al., "A Neuromorphic Visual System Using RRAM Synaptic Devices with Sub-pJ Energy and Tolerance to Variability: Experimental Characterization and Large-Scale Modeling," *IEDM Technical Digest*, p. 10.4.1, 2012.
8. Park, S. et al., "RRAM-Based Synapse for Neuromorphic System with Pattern Recognition Function," *IEDM Technical Digest*, p. 10.2.1, 2012.
9. Prezioso, M., F. Merrikh-Bayat, B.D. Hoskins, G.C. Adam, K.K., Likharev, "Training and Operation of an Integrated Neuromorphic Network Based on Metal-Oxide Memristors," *Nature*, Vol. 521, pp. 61-64, May 2015.
10. Prezioso, M., I. Kataeva, F. Merrikh-Bayat, B. Hoskins, G. Adam, T. Sota, K. Likharev, and D. Strukov, "Modeling and Implementation of Firing-Rate Neuromorphic-Network Classifiers with Bilayer Pt/Al_2O_3/TiO_{2-x}/Pt memristors", accepted to *IEDM'15*, Dec. 2015.
11. Alibart, F., L. Gao, B. Hoskins and D.B. Strukov, "High-Precision Tuning of State for Memristive Devices by Adaptable Variation-Tolerant Algorithm," *Nanotechnology*, Vol. 23, art. 075201, 2012.
12. Alibart, F., A. Zamanidoost, and D.B. Strukov, "Pattern Classification by Memristive Crossbar Circuits with Ex-situ and In-situ Training," *Nature Communications*, Vol. 4, p. 2072, 2013.
13. Schiffmann, W., M. Joost, and R. Werner, "Optimization of the Backpropagation Algorithm for Training Multilayer Perceptrons," Technical Report, University of Koblenz, 1994.
14. Merrikh Bayat, F., B. Hoskins, and D.B. Strukov, "Phenomenological Modeling of Memristive Devices," *Applied Physics A*, Vol. 118, pp. 770-786, 2015.
15. Ciresan, D., U. Meier, and J. Schmidhuber, "Multi-Column Deep Neural Networks for Image Classification," *Proceedings of CVPR'12*, Providence, RI, June 2012, pp. 3642-3649.
16. Kataeva, I., F. Merrikh Bayat, E. Zamanidoost, and D.B. Strukov, "Efficient Training Algorithms for Neural Networks Based on Memristive Crossbar Circuits", *Proceedings of IJCNN'15*, Killarney, Ireland, July 2015, pp. 1-8.
17. Kataeva, I., T. Sota, T. Rojanaarpa, F. Merrikh-Bayat and D. Strukov, "Efficient Hardware-Compatible Training Algorithms for Neural Networks Based on Memristive Circuits", in preparation, Jan. 2016.
18. Govoreanu, B. et al., "Vacancy-Modulated Conductive Oxide Resistive RAM (VMCORRAM)," *IEDM Technical Digest*, 10.2, p. 10.2.1, 2013.